A population of neurons selective for human voice in the monkey brain

Margherita Giamundo, Regis Trapeau, Etienne Thoret, Luc Renaud, Simon Nougaret, Thomas Brochier, Pascal Belin

PNAS RESEARCH ARTICLE | NEUROSCIENCE

A population of neurons selective for human voice in the monkey brain

Margherita Giamundo[a,b,1], Regis Trapeau[a], Etienne Thoret[a,b], Luc Renaud[a], Simon Nougaret[a], Thomas G. Brochier[a], and Pascal Belin[a,b,1]

Edited by Stephen G. Lomber, McGill University, Montreal, QC, Canada; received March 22, 2024; accepted April 25, 2024 by Editorial Board Member Michael E. Goldberg

Many animals can extract useful information from the vocalizations of other species. Neuroimaging studies have evidenced areas sensitive to conspecific vocalizations in the cerebral cortex of primates, but how these areas process heterospecific vocalizations remains unclear. Using fMRI-guided electrophysiology, we recorded the spiking activity of individual neurons in the anterior temporal voice patches of two macaques while they listened to complex sounds including vocalizations from several species. In addition to cells selective for conspecific macaque vocalizations, we identified an unsuspected subpopulation of neurons with strong selectivity for human voice, not merely explained by spectral or temporal structure of the sounds. The auditory representational geometry implemented by these neurons was strongly related to that measured in the human voice areas with neuroimaging and only weakly to low-level acoustical structure. These findings provide new insights into the neural mechanisms involved in auditory expertise and the evolution of communication systems in primates.

voice processing | temporal voice areas | fMRI-guided electrophysiology

Significance

Humans and other animals have specialized brain regions dedicated to processing the voice of their conspecifics. Here, we show the involvement of neurons in these regions also in processing the voices of other species with which they have daily exposure. These findings shed light on the neural mechanisms underlying interspecies communication, such as between predators and prey in the wild or, in the context of our everyday life, in the interactions with our domestic pets.

Temporal voice areas (TVAs) are regions of the cerebral cortex that are involved in the processing of voice information. They have been observed in humans, rhesus macaques, and common marmosets (1–5), and analogous regions have been discovered in dogs and cats (6, 7). The TVAs not only exhibit increased activity in response to conspecific vocalizations (CVs) but also categorize CVs apart from other sounds in a functionally homologous manner in humans and macaques (3). They are thought to be part of an evolutionary-ancient "voice patch system" devoted to processing behaviorally relevant information from voices (8), analogous to the "face patch system" of the primate visual cortex (9).

Yet many animals also extract valuable information from the vocalizations of other species, such as predators, prey, pets, or, in the case of laboratory rhesus macaques, humans (10). How the brain processes such information remains unclear. A strict evolutionary account would predict that the TVAs of a given species represent only the vocalizations of living beings with which that species has coevolved, including conspecifics as well as long-standing predators and prey. However, expertise acquired during a lifetime could also affect these representations, as suggested by data showing a strong modulatory effect of early experience with faces on macaque face patch activity (11).

To investigate this issue, we used fMRI-guided electrophysiology in two macaque monkeys to localize their individual TVAs before surgically implanting high-density chronic electrode arrays at these locations. We recorded the spiking activity of individual voice-patch neurons in response to a wide number of complex sounds including macaque and human vocalizations and investigated the category selectivity of these neurons. After confirming the existence of neurons selective for macaque vocalizations in the TVAs (12), we identified a population of neurons selective for, and representing, human voice.

Results

Not only Macaque-Selective, but also Human-Selective Neurons in the aTVA. We conducted electrophysiological recordings in the fMRI-localized anterior TVA (aTVA) of two female rhesus monkeys, M1 and M2 (Fig. 1*A*) (2, 3). The aTVA was targeted because it was the most prominent voice patch in the two monkeys and because of its functional homology to the human aTVA in categorizing CVs apart from other sounds (3). M1 was implanted with two 32-channel Utah arrays in the aTVA located on the rostral superior temporal gyrus (rSTG) of the right hemisphere; M2 was implanted with one 32-channel Utah array in the aTVA located on the rSTG close to the upper bank of the superior temporal sulcus (STS) in the left hemisphere. The targeted recording areas resided in hierarchically high-level auditory cortex near the temporal pole and showed no clear tonotopic organization (*SI Appendix*, Fig. S1*A*).

Author affiliations: [a]Institut de Neurosciences de la Timone, Aix-Marseille Université, UMR 7289 CNRS, Marseille 13005, France; and [b]Institute of Language Communication and the Brain, Marseille 13000, France

Author contributions: M.G., R.T., and P.B. conceptualized research; M.G., R.T., E.T., T.G.B., and P.B. developed methodology; L.R. and S.N. conducted surgical procedures; M.G. analyzed data; M.G. and P.B. produced visualizations; M.G. and P.B., administered the project; T.G.B. and P.B. supervised research; and M.G., R.T., E.T., L.R., S.N., T.G.B., and P.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. S.G.L. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: margherita.giamundo@univ-amu.fr or pascal.belin@univ-amu.fr.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2405588121/-/DCSupplemental.

Published June 11, 2024.

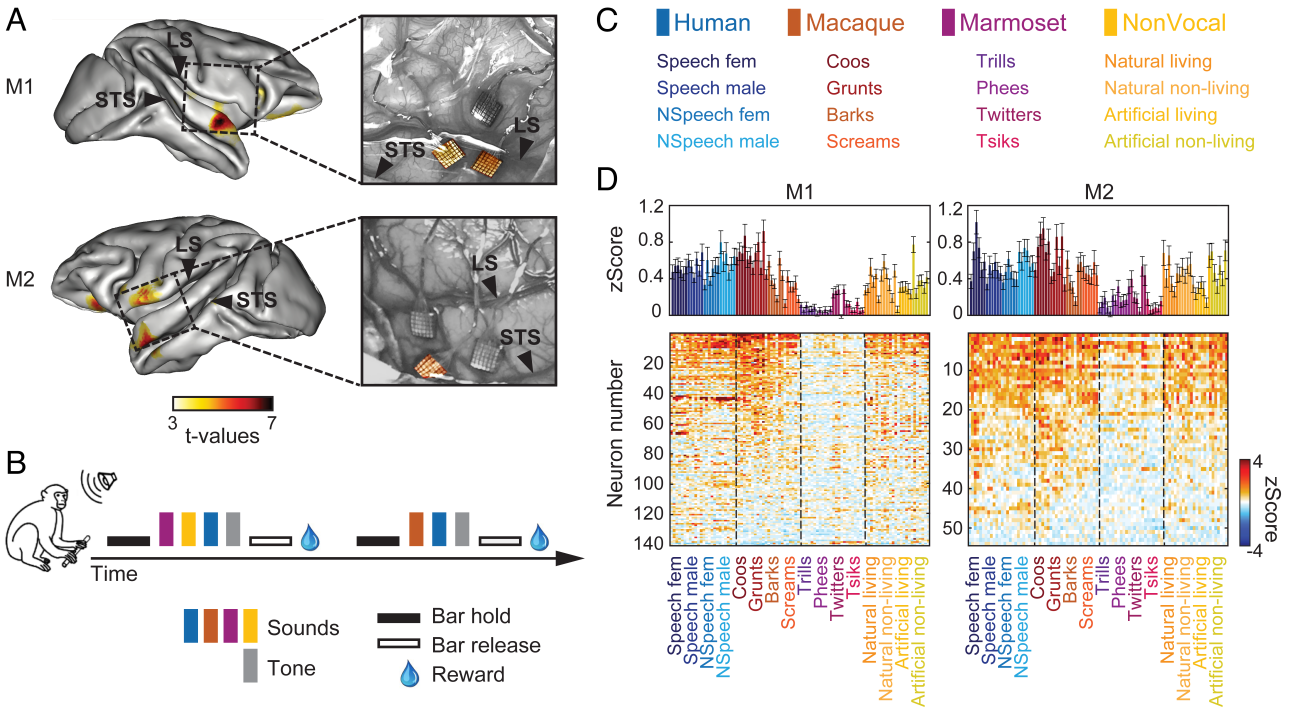

Fig. 1. Electrophysiological recordings of spiking activity from aTVA neurons. (*A*) Implantation sites of high-density chronic electrode Utah Arrays in monkeys M1 and M2. The color scale indicates t-values of fMRI contrast between macaque vocalizations vs. nonvocal sounds. Utah arrays analyzed here (highlighted in the pictures of the cortical surface during surgery, *Right Insets*) were implanted in cortical areas close to the fMRI-identified aTVA peaks. STS: superior temporal sulcus. LS: lateral sulcus. (*B*) Pure tone detection task. Monkeys were trained to release a bar when a pure tone was randomly presented among other stimuli to obtain the reward. (*C*) Stimuli consisted of 96 complex natural sounds divided into four main categories, each divided into four subcategories. (*D*) Neurons responsiveness (z-scores computed during the maximum response time; see *Materials and Methods*) to the 96 sounds for M1 (*Left*) and M2 (*Right*). *Bottom*: Each line represents the response of one neuron; neurons are sorted depending on the strength of macaque vocalization–evoked response. *Top*: Stimulus-specific average population response (mean ± SE).

Recordings of spiking activity took place while the monkeys performed active detection of a pure tone interspersed among a set of 96 auditory stimuli (pure tone detection task; Fig. 1*B*). The task was introduced to maintain the attention of the monkeys to the auditory stimulation. The set of stimuli, previously presented to the monkeys and human participants during fMRI scanning (3), consisted of brief complex sounds sampled from 4 large categories: macaque (conspecific) vocalizations (N = 24), human (behaviorally relevant heterospecific) voices (N = 24), marmoset (unknown heterospecific) vocalizations (N = 24), and nonvocal (control) sounds (N = 24). *SI Appendix*, Fig. S1 *B* and *C*, shows example spectrograms and the long-term average spectrum of each category. Each category was also divided into four subcategories (Fig. 1*C*). We focused our analyses on a total of 194 well-isolated auditory responsive neurons—140 out of 260 neurons in M1 and 54 out of 93 neurons in M2—i.e., neurons responding to at least one sound (Fig. 1*D*; see *Materials and Methods*).

Populations of auditory responsive neurons in aTVA were able to discriminate between the four sound categories (max correlation coefficient classifier, 100 ms time windows, 10 ms time bins; see *Materials and Methods*) about a few milliseconds after the stimulus onset (*SI Appendix*, Fig. S2*A*). The peak of the discrimination occurred around 100 ms after the stimulus onset (M1: 130 ms; M2: 100 ms), consistent with the latency of peak of activity in the majority of neurons (*SI Appendix*, Fig. S2*B*).

Fig. 2*A* shows three example neurons that were more responsive to macaque vocalizations (*Left*), human voices (*Central*), or both (*Right*) compared to the other sound categories (see also *SI Appendix*, Fig. S3 for example neurons from each array and monkey). This pattern of higher responsiveness to macaque and/or human voices was representative of the aTVA neuronal population. Indeed, we found proportions of categorical responsiveness very similar between aTVAs of M1 and M2. Neurons responding maximally (preferred category; see *Materials and Methods*) to macaque vocalizations or to human voices were more than two times (in M1) or three times (in M2) as numerous as those responding maximally to marmoset or to nonvocal sounds (Fig. 2*B*; Chi-squared test comparing to a uniform distribution of 20%, M1: $\chi 2 = 59.64$, df = 4, $P = 3.448e\text{-}12$; M2: $\chi 2 = 26.56$, df = 4, $P = 2.445e\text{-}05$). Furthermore, the average population spiking activity was significantly higher for macaque and human voices compared to marmoset vocalizations and to nonvocal sounds (Fig. 2*C*; two-way ANOVA with category and monkey as factors; category effect: $F_{3,768} = 29.95$, $P = 0$; multiple comparison post hoc tests: human vs. macaque $P = 0.99$, for all other comparisons $P < 0.01$; no monkey effect: $F_{1,768} = 2.53$, $P = 0.11$; no interaction category x monkey: $F_{3,768} = 0.04$, $P = 0.99$).

To further investigate the preference of aTVA neurons for macaque and human vocalizations, we computed two Voice Selectivity Indices (VSIs) for each neuron (in line with previous works in the voice (12) and face (13) domains; see *Materials and Methods*): one contrasting macaque vocalizations with the nonvocal category and the second contrasting human voices with the nonvocal category. Fig. 2*D* shows the human (y-axis) and macaque (x-axis) VSIs of all neurons from the two monkeys. The distributions of human and macaque VSIs were significantly correlated (Spearman correlation, M1: rho = 0.3135, $P = 1.6182e\text{-}04$; M2: rho = 0.4700, $P = 3.3640e\text{-}04$; see also *SI Appendix*, Fig. S4*A*), with a majority of neurons more responsive to vocalizations compared to nonvocal sounds (M1: 69% of neurons with a human VSI > 0 corresponding to higher activity for human vs. nonvocal, 62% with a macaque VSI > 0; M2: 74% with a human VSI > 0, 69% with a macaque VSI > 0). However, while around 20% (39 of 194) of neurons were selective (VSI > 0.33, corresponding to a neuronal response to vocalizations at least two times

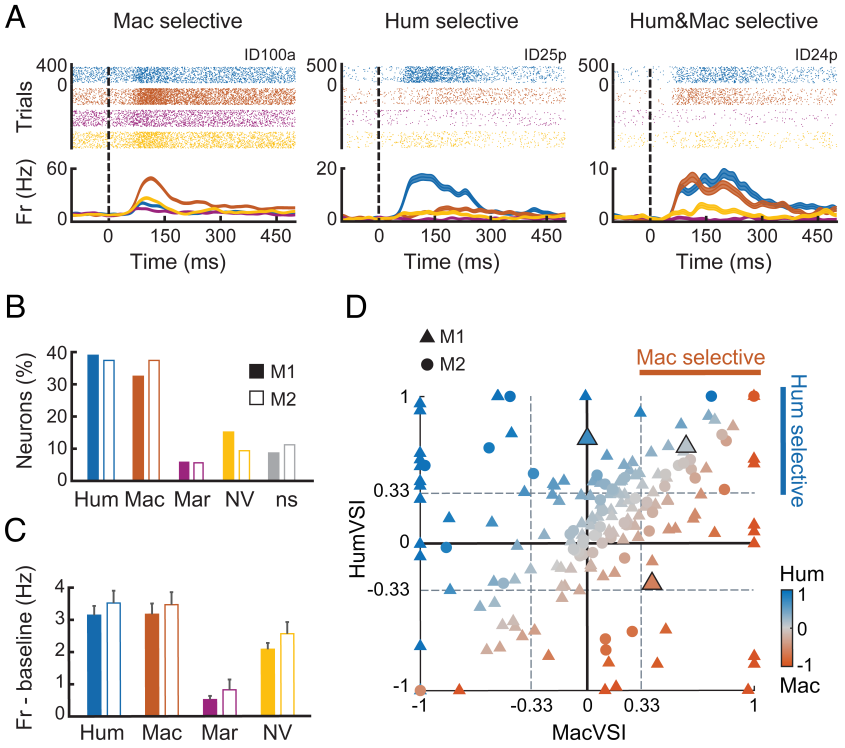

Fig. 2. aTVA neurons are selective for human and/or macaque vocalizations. (*A*) Example aTVA neurons with strongest response to macaque vocalizations (*Left*), to human voices (*Middle*), or to both (*Right*). Each panel shows raster plots (*Top*) and spike densities (*Bottom*) per sound category (human: blue; macaque: orange; marmoset: magenta; nonvocal: yellow), relative to sound onset. Shaded areas indicate ± SE. Mac selective: macaque-selective neuron (cf. main text); Hum selective: human-selective neuron; Hum&Mac selective: neuron selective for both human and macaque vocalizations. (*B*) Proportion of aTVA neurons in M1 (N = 140) and M2 (N = 54) showing strongest response to each sound category (preferred category; see *Materials and Methods*) calculated by performing a one-way ANOVA with category as factor (*P*-value threshold = 0.05). ns: no significant preference. Hum: human voices; Mac: macaque vocalizations; Mar: marmoset vocalizations; NV: nonvocal sounds. (*C*) Average population firing rate (mean ± SE) relative to prestimulus baseline for each sound category and monkey. (*D*) Distribution of human and macaque Voice Selectivity Indices (VSIs) across all neurons (N = 194) of M1 (triangular symbols) and M2 (circular symbols). The largest symbols represent VSIs of example neurons in (*A*). One neuron is considered selective for voices if VSI > 0.33, i.e., its spiking rate is more than double that for the nonvocal category. The color bar indicates the SI contrasting human vs. macaque vocalizations (*Materials and Methods*).

stronger than to nonvocal sounds; see refs. 12 and 13) for both human and macaque vocalizations, many neurons were selective exclusively for human voices (20%, 39 of 194) or for macaque vocalizations (12%, 23 of 194). In comparison, 7% (15 of 194) of neurons were exclusively selective (i.e., HumVSI and MacVSI < -0.33) to nonvocal sounds. The anatomical localization of the human- and macaque-selective neurons showed no clear topographic organization (*SI Appendix*, Fig. S4*B*). Subsequent analyses were focused on the two subpopulations of macaque- (N = 23) or human-selective neurons (N = 39) (but see *SI Appendix*, Fig. S5*A* for analysis of neurons selective for both human and macaque vocalizations).

Human-Selective Neurons Represent Human Voices Apart from Other Sounds. To confirm the specialization of the human- and macaque-selective subpopulations for one particular category of voices, a machine-learning classifier (max correlation coefficient, see *Materials and Methods*) was trained to discriminate between human or macaque vocalizations from the other sound categories based on population spiking activity (100 ms time windows every 10 ms time bins). For both subpopulations of neurons (Fig. 3*A*), accuracy was significantly above chance level from about sound onset throughout sound duration (all *P*'s < 0.0002, permutation tests; see *Materials and Methods*). However, in the human-selective subpopulation, classification accuracy was higher for human voices than for macaque vocalizations (Fig. 3 *A*, *Top* panel; from 150 to 470 ms after onset all η values < 0.05; see *Materials and Methods*), whereas accuracy was higher for macaque than for human vocalizations in the macaque-selective subpopulation (*Bottom* panel; from 90 to 140 ms after onset all η values < 0.05). Interestingly, while classification accuracy profiles for macaque vocalizations vs. other sounds were similar for the two subpopulations of neurons (Fig. 3*A*, red curves), classification of human voice vs. other sounds markedly differed between subpopulations (blue curves).

For a finer-grained understanding of sound representation in these neurons, we used representational similarity analysis (RSA) (14). RSA is complementary to the above classification analyses by examining continuous distances between stimulus representations rather than categorical boundaries. For each subpopulation (macaque-selective and human-selective) and each 10 ms bin across stimulus presentation time, we built a 16 × 16 Neuronal representational dissimilarity matrix (RDM) capturing the pattern of pairwise dissimilarity population spiking activity (Euclidean distance in multineuron space) to each pair of the 16 subcategories in a 100 ms window centered on the bin (Fig. 3 *B*, *Top*).

We compared the obtained neuronal RDM time series to four categorical model RDMs (Fig. 3 *B*, *Bottom*), representing the theoretical patterns of pairwise dissimilarities in our stimulus set under four separate assumptions of ideal categorical distinction (3): 1- a human model in which human voices are categorized apart from all other sounds, with no dissimilarity between neuronal responses to pairs of human voices or to pairs of the other sounds, but maximal dissimilarity between responses to human voices vs. other sounds; 2- a macaque model categorizing macaque vocalizations apart from other sounds; 3- a marmoset model categorizing marmoset vocalizations apart from other sounds; and 4- a nonvocal model categorizing nonvocal sounds apart from vocalizations of all species.

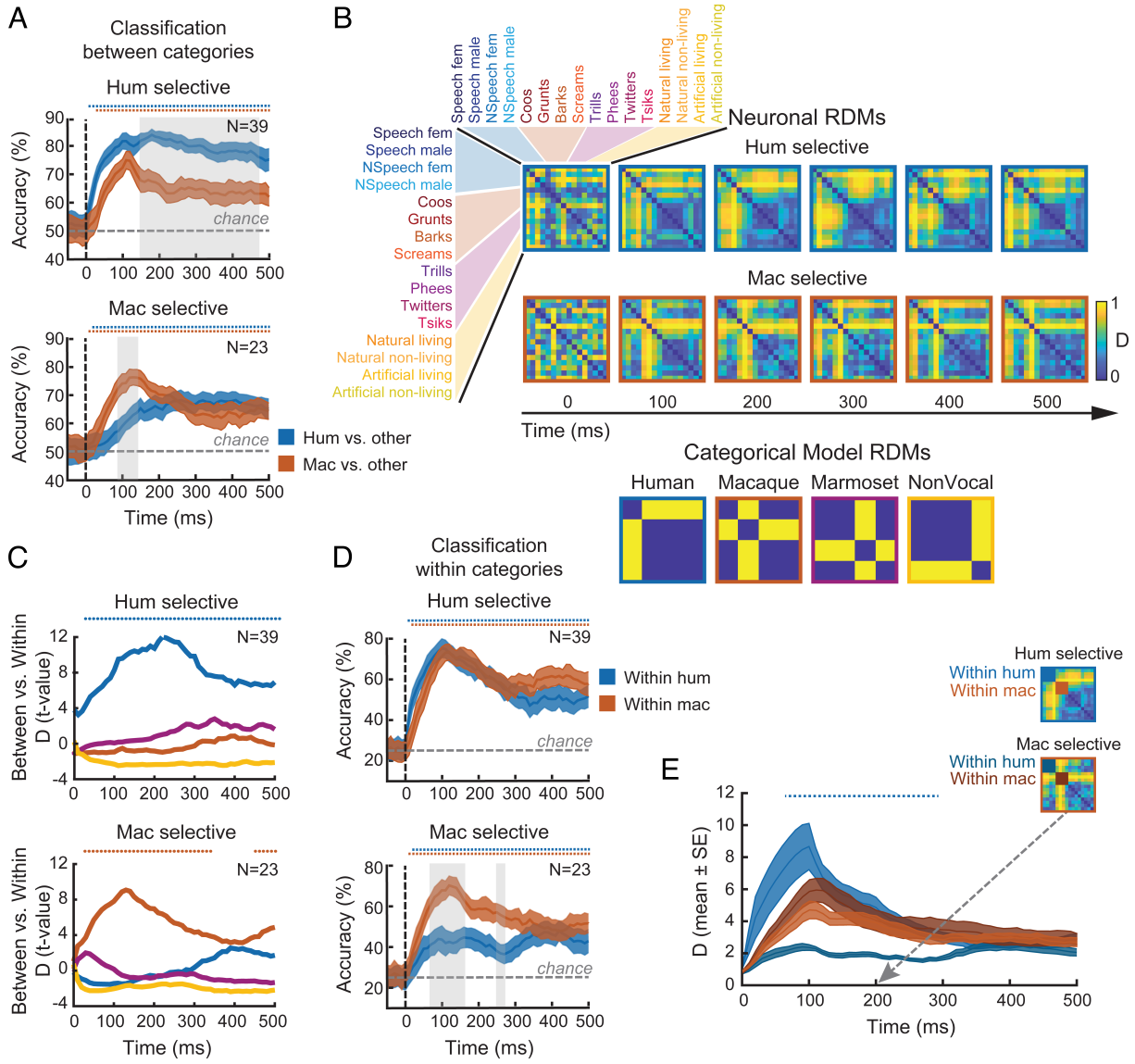

Fig. 3. Representation of human and macaque vocalizations in human- and macaque-selective neurons. (*A*) Time-resolved classification accuracy (mean ± SE) of linear classifiers trained to discriminate stimuli between human (blue curves) or macaque (red curves) category vs. the other sound categories (chance level: 50%) based on spiking activity of either human- (*Top*) or macaque- (*Bottom*) selective neurons across the two monkeys. Colored dots above indicate time bins of significantly above-chance classification accuracy (permutation tests, $P < 0.0002$; see *Materials and Methods*). The gray shaded areas indicate time bins of significant difference between classification accuracies for human vs. macaque vocalizations ($\eta < 0.05$, see *Materials and Methods*). (*B*) *Top* panels: Neuronal RDMs capturing spiking differences across all pairs of the 16 sound subcategories as a function of postonset time for the human-selective (*Top*) and macaque-selective (*Bottom*) neuronal subpopulations. The color scale indicates normalized pairwise distance rankings. *Bottom* panels: Categorical model RDMs representing ideal categorical distinctions between human voices vs. all other sounds, macaque vocalizations vs. all other sounds, marmoset vocalizations vs. all other sounds, or nonvocal sounds vs. all vocalizations. (*C*) Time course of associations between human-selective (*Top*) and macaque-selective (*Bottom*) neuronal RDMs and model RDMs. Color code as in (*B*); human: blue, macaque: orange, marmoset: magenta, nonvocal: yellow. Colored dots indicate time bins of statistically significant association (between/within bootstrapped two-sample *t* tests, Bonferroni-corrected threshold of $P = 2.451e\text{-}04$). Neuronal RDMs in both subpopulations only show association with the corresponding model RDM. (*D*) Accuracy (mean ± SE) in the classification of stimuli within the human category (i.e., between the four different human subcategories; blue curve) or within the macaque category (i.e., between the four macaque subcategories; red curve; chance level: 25%) for human- (*Top*) or macaque- (*Bottom*) selective neurons across the two monkeys. Colored dots indicate time bins with significantly above-chance accuracy (permutation tests, $P < 0.0002$). The gray shaded areas indicate time bins with significant difference between classification accuracies for human vs. macaque vocalizations ($\eta < 0.05$). Note how classification of human voice subcategories differs across the two subpopulations (blue curves), in contrast to macaque subcategories (red curves). (*E*) Dissimilarity values (*D*; mean ± SE) of the portions of RDMs within the human (blue curves) or within the macaque (red curves) categories, for human- (N = 39) and macaque-selective (N = 23) neurons. Colored dots indicate time bins with a significant difference between human- and macaque-selective neurons (Mann–Whitney–Wilcoxon test, $P < 0.05$). While dissimilarities between macaque subcategories are similar across the two populations, they markedly differ for human voice subcategories (blue dots).

In the human-selective subpopulation, associations between model and neuronal RDMs only reached significance for the human model (Fig. 3 *C*, *Top* panel; bootstrapped two-sample *t* tests between distance values in the between vs. the within parts of the neuronal RDMs predicted by the model, 10,000 iterations, Bonferroni-corrected threshold of $P = 2.451e\text{-}04$), from around 30 ms poststimulus onset throughout sound duration. Similarly, in the macaque-selective subpopulation (Fig. 3 *C*, *Bottom*), only the association with the macaque model was significant, from around 30 ms poststimulus onset. Thus, the firing rate of human- and macaque-selective neurons allows for better classification of sounds from their corresponding category and reflects a representational geometry that categorizes these sounds apart from all others, particularly between 100 and 200 ms after sound onset.

Human-Selective Neurons Represent Human Voice Subcategories More Finely than Macaque-Selective Neurons. To explore the informational content of vocalizations firing rate contains beyond mere categorical separation, we trained a linear classifier to categorize stimuli between the four macaque subcategories (corresponding to different call types within the macaque category) or between the four human subcategories (corresponding to different genders or speech content within the human category). Indeed, it is possible that although the human-selective subpopulation can discriminate human voices from the other sound categories, it cannot discriminate between different types of human voices, i.e., voices that are not from conspecifics. In other words, how refined is the representation of human voices by this subpopulation of neurons? As observed above for category decoding, decoding of macaque vocalization subcategories showed similar profiles for the two subpopulations (Fig. 3*D*, red curves). However, decoding of human voice subcategories differed markedly (Fig. 3*D*, blue curves): While accuracy matched that for macaque subcategories in the human-selective subpopulation, it was significantly lower in the macaque-selective subpopulation (between 70 to 160 ms and 250 to 270 ms after stimulus onset, all η values < 0.05). These results were consistent also when more restrictive parameters were used to select the human- and macaque-selective neurons (*SI Appendix*, Fig. S5 *B* and *C*).

To confirm this finding, we directly compared the neuronal RDMs from the human-selective subpopulation with those from the macaque-selective subpopulation, focusing on the dissimilarities within the human or the macaque categories. We found no difference within macaque vocalization category between the human- vs. macaque-selective subpopulations (Fig. 3*E*, red curves; Mann–Whitney–Wilcoxon test, all *P*s > 0.05, ns). In contrast, the human-selective subpopulation showed markedly larger dissimilarities within human voices category than the macaque-selective subpopulation from about 40 ms until 290 ms poststimulus onset (*P*s < 0.05; Fig. 3*E*, blue curves). Thus, the human-selective neurons did not differ from the macaque-selective neurons in the representation of the macaque vocalization subcategories. However, they showed a fine-grained representation of human voice subcategories, not present in the macaque-selective neurons.

Similar Patterns of Sound Representation for Human-Selective Neurons and Human fMRI Data Clustering Apart from Acoustical Representations. How similar are sound representations by this human-selective macaque's neuronal subpopulation to those by the human aTVA or to acoustical representations? We addressed this question, comparing neuronal RDMs (computed at the time of peak association with the respective model RDM) with fMRI and acoustical RDMs, as well as with the categorical model RDMs (Fig. 4*A*; see *Materials and Methods*). Because data on single-neuron spiking activity in the human TVAs are not currently available, we used fMRI RDMs (14) obtained from fMRI measures of the neural activity in the human and macaque aTVAs (3) in response to the same set of stimuli (Fig. 4*A*; see *Materials and Methods*). Acoustical RDMs were computed from acoustical measures of the stimuli in order to consider a possible explanation of dissimilarities based on the acoustical differences between sound categories (*SI Appendix*, Fig. S1*C*). Three acoustical measures of the stimuli were considered: loudness, spectral center of gravity (SCG), and pitch (Fig. 4*A*; see *Materials and Methods*).

Human- and macaque-selective neuronal RDMs weakly correlated (bootstrapped Spearman's rho = 0.1043, bootstrapped *P* = 0.27, above Bonferroni-corrected threshold of *P* = 0.05/17 = 0.0029, n.s.; Fig. 4*B*). In contrast, the two neuronal RDMs strongly correlated with the corresponding model RDMs (human-selective neuronal RDM with human model: rho = 0.7351, *P* = 0;

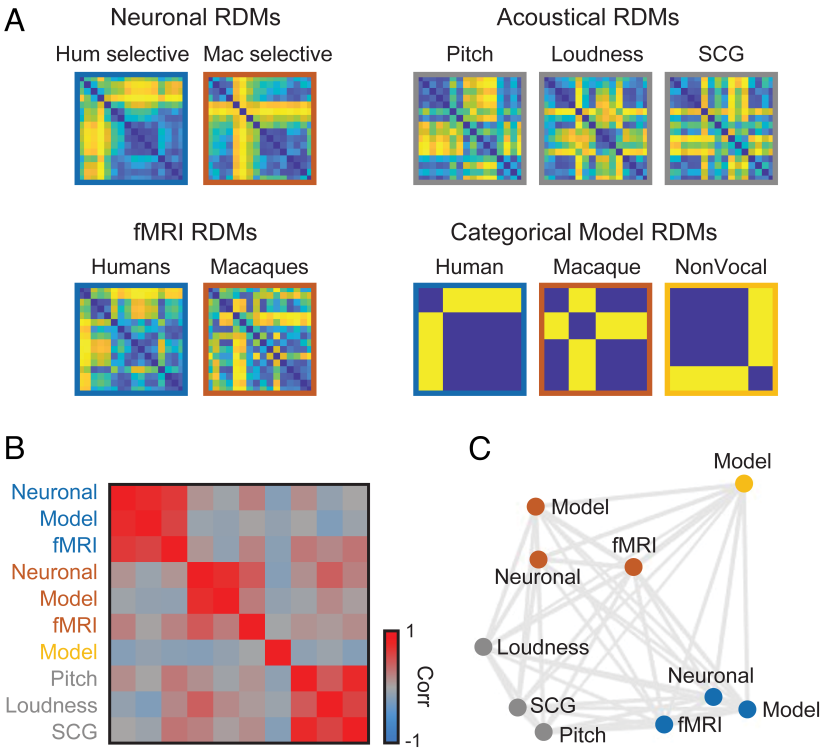

Fig. 4. Comparison of stimulus subcategory representations provided by neuronal measures, fMRI activations, acoustical descriptors, and categorical models. (*A*) The different RDMs compared: neuronal RDMs based on spiking activity for the two subpopulations, computed at the time of peak association with the respective model RDM; fMRI RDMs based on aTVA activity measured using fMRI in humans and in macaques; acoustical RDMs based on three acoustical features; and categorical model RDMs. (*B*) Correlation matrix (Spearman rank correlation) of the 10 RDMs. (*C*) 2D representation of correlations between RDMs via multidimensional scaling. Large distances indicate low correlations. Note the distance between human- (blue) and macaque-selective (orange) neuronal RDMs and their clustering with the corresponding model and fMRI RDMs (symbols with the same color code), but not with acoustical RDMs (gray symbols).

macaque-selective neuronal RDM with macaque model: rho = 0.7307, P = 0), and the corresponding fMRI RDMs (human-selective neuronal RDM with human fMRI RDM: rho = 0.6494, P = 0; macaque-selective neuronal RDM with macaque fMRI RDM: rho = 0.4574, P = 0). No significant association was found between neuronal RDMs and their mismatched model and fMRI RDMs (all rhos > −0.193 and <0.2329, all P's below Bonferroni-corrected threshold of P = 0.05/17 = 0.0029, n.s.). We also found little correlation between the neuronal RDMs and the acoustical RDMs, only reaching significance for the comparison between macaque-selective neuronal RDM and loudness (rho = 0.4183, P = 0; all other comparisons: rhos < 0.2166, all P's < 0.0029, n.s.; Fig. 4*B*).

This pattern of dissimilarities is illustrated in Fig. 4*C* as a multidimensional scaling representation (15) of associations between the different RDM types (neuronal, model, fMRI, and acoustical) in which distances between symbols are inversely related to correlation strength. From this geometrical representation of dissimilarities, the neuronal RDMs tightly clustered with their corresponding model and fMRI RDMs (Fig. 4*C*, symbols with the same color code), but not with acoustical RDMs (gray symbols).

Human-Voice Selectivity Is Not Explained by Spectral or Temporal Tuning. An alternative explanation to our findings other than in terms of voice selectivity could be in terms of tuning to low-level acoustical features, such as a specific spectral shape or temporal envelope. To test that hypothesis, we acquired additional data in M1 from four sessions in which we presented the 96 original stimuli of the former main experiment along with two acoustical controls for each stimulus: a spectrally matched (SM) and a temporally matched (TM) controls (Fig. 5*A*). Five band-passed noise stimuli (*Materials and Methods*) were also included in each session in order to relate voice selectivity to frequency tuning at the single unit level. Spectrally and temporally matched stimuli corresponded to a transformation of the original sounds in which the temporal and spectral contents are kept constant, respectively (*Materials and Methods*). This allowed to test the encoding of each of these low-level acoustic features respectively. Indeed, if the apparent selectivity reflects tuning to specific low-level acoustics, then similar selectivity should be observed for stimuli having the same spectral or temporal distribution as the original stimuli.

We were able to extract 38 auditory responsive neurons in these additional recordings that confirmed our initial observations despite the 24-mo interval: We found a sizeable proportion of units (13 on

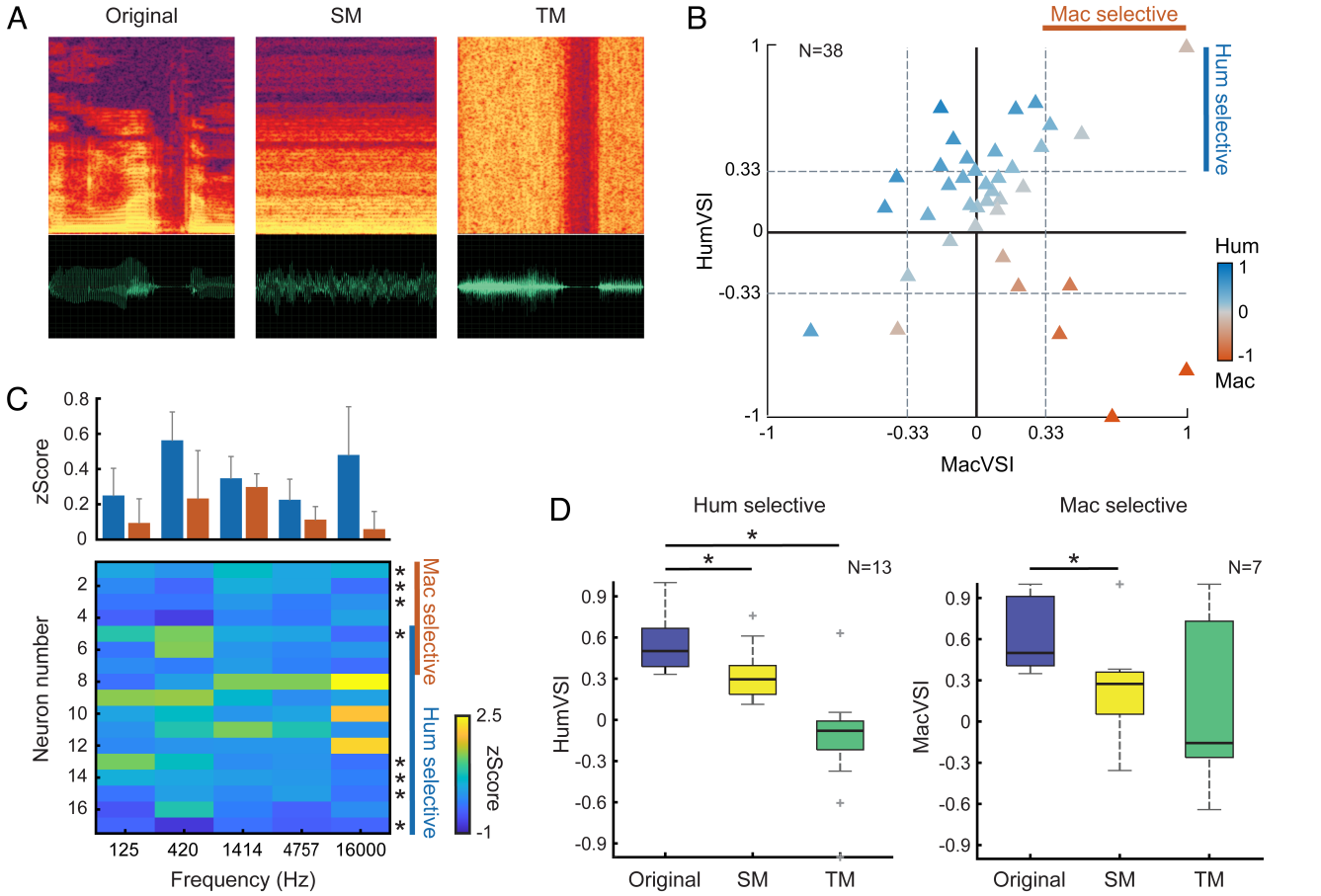

Fig. 5. Lack of association between voice selectivity and spectral or temporal structure. (*A*) Spectrograms (*Top*) and waveforms (*Bottom*) of an example stimulus (human speech) from the original stimulus set and its spectrally matched (SM) and temporally matched (TM) controls. (*B*) Distribution of human and macaque Voice Selectivity Indices (VSIs) across all auditory responsive neurons of M1 recorded in control sessions. A neuron is considered selective for VSI > 0.33. The color bar indicates the SI contrasting human vs. macaque vocalizations. (*C*) Frequency tuning of the voice-selective neurons. For each neuron (N = 17; lines), the z-scored FR was computed during the maximum response time respect to the baseline, in response to the 5 band-passed noise stimuli (columns). The plot on the top shows the average (± SE) across human-selective (N = 13) and macaque-selective (N = 7) neurons. *: Neurons with different (one-way ANOVA, P < 0.05) tuning in response to the 5 band-passed noise stimuli. No significant difference was found at the population level (Kruskal–Wallis test, P-value threshold = 0.05). (*D*) Boxplots of the human (*Left*) and macaque (*Right*) VSIs in their corresponding selective subpopulations for the three conditions of stimuli. VSIs are significantly reduced (paired-samples Wilcoxon signed-rank test, P < 0.05; *) for the SM stimuli in both populations and for the TM stimuli in the human-selective neurons.

38) with a human VSI above 0.33, and fewer neurons (7 on 38) selective for macaque vocalizations, of which only three neurons selective for both (Fig. 5*B*).

First, we examined the frequency tuning of each selective neuron (Fig. 5*C*). There were some neurons with a different tuning in response to the five band-passed noise stimuli (human-selective: 5 of 13 neurons; macaque-selective: 4 of 7 neurons; one-way ANOVA, $P < 0.05$; Fig. 5 *C*, *Bottom*), but without any consistent frequency tuning at the population level (Kruskal–Wallis test, human-selective: $\chi^2 = 4.78$, df = 4, $P = 0.3104$; macaque-selective: $\chi^2 = 3.65$, df = 4, $P = 0.4558$; Fig. 5 *C*, *Top* panel). We then examined VSIs of those selective neurons for the control stimuli: Compared to the original, VSIs significantly decreased for the SM controls in both subpopulations (paired-samples Wilcoxon signed-rank test; human-selective: $P = 0.0266$; macaque-selective: $P = 0.0312$) and decreased for the TM stimuli in the human-selective subpopulation (human-selective: $P = 0.0002$; macaque-selective: $P = 0.0781$; Fig. 5*D*). These results indicate that the selectivity of aTVA neurons for human voices (and/or for macaque vocalizations) does not reflect a simple tuning associated to low-level acoustical features.

Discussion

In the present study, we have identified a subset of neurons in the TVAs of laboratory macaques that are specifically involved in representing important aspects of human voices. Initially, we confirmed the presence of conspecific voice cells in the anterior voice patch, which are neurons that selectively respond to macaque vocalizations—an observation previously reported by only a single research group (12, 16). Subsequently, we identified another group of neurons that selectively respond to human voices but not macaque vocalizations. The firing rate of these human-selective neurons allowed for more accurate classification of human voices compared to other sounds, in contrast to macaque vocalizations. Furthermore, the classification between subcategories of human voices (such as male/female and speech/nonspeech) by these neurons was comparable to the classification of macaque vocalization subcategories (coo/scream/grunt/bark). Moreover, the representation patterns of these human-selective neurons were more similar to those observed in fMRI scans of human brains rather than macaque brains and only weakly correlated with acoustical RDMs. This finding is consistent with previous fMRI data indicating that while RDMs from primary auditory cortex correlate strongly with acoustical RDMs in both humans and these two macaques, it is not the case for their higher-level aTVAs that implement a more categorical representation (3). This, as well as the reduced VSIs observed for both the SM and the TM controls, and the lack of consistent frequency tuning in these neurons, reinforces the notion that their human-voice selectivity is not merely reflecting low-level acoustical features but a more abstract representation of this particular sound category.

The presence of macaque neurons that specifically respond to human voices is inconsistent with a pure evolutionary explanation of voice selectivity. Given that wild macaques have not had significant exposure to the human voice during the past millions of years, it is difficult to understand the selective pressures that would lead to the development of neurons specifically tuned to human vocalizations. A more parsimonious explanation of our findings is in terms of expertise: The macaques involved in our experiments have been raised in captivity and have been exposed to human vocal sounds since birth. These sounds, that announce the arrival of their assigned experimenter or of the veterinarian, are highly relevant to them. The almost total lack of responsiveness to unknown marmoset vocalizations seems to reinforce this hypothesis, although the markedly different frequency structure of marmoset calls with an energy peak between 6 and 8 kHz rather than between 0 and 2 kHz (*SI Appendix*, Fig. S1) certainly is a key contributing factor. Thus, the selectivity of these neurons to human voice and their capability of processing information from such stimuli are likely the result of lifelong daily exposure.

Remarkably, while the firing rate of both the human- and macaque-selective populations yielded comparable classification of macaque vocalizations, it was specifically in the classification of human voices that the human-selective population exhibited a clear advantage. This, and the apparent intermingling of these two populations, suggests that the human-selective neurons may have leveraged preexisting mechanisms primarily used for processing conspecific vocalizations to decode information in the relevant sound category of human voice, aided in this by the close similarity in the human and macaque vocal apparatuses. This utilization of preexisting machinery for a novel purpose is reminiscent of similar phenomena observed in the human visual cortex (17), where areas specialized for face recognition are recruited by experts in their respective domains (18).

Overall, these findings not only enhance our understanding of the neural processes involved in voice processing and the evolution of primate communication systems but also shed light on the neuronal mechanisms potentially underlying other forms of auditory expertise, such as musical expertise, in the human brain.

Materials and Methods

Experimental Design.

Subjects. Data from the main experiment were recorded from two female rhesus monkeys (Macaca mulatta, M1 and M2, aged 7 and 8 y respectively, and weighing between 5 and 6 kg). In the control experiment, data were collected from only M1. Animal care, housing, and experimental procedures were in compliance with the NIH's Guide for the Care And Use of Laboratory Animals and approved by the Ethical Board of Institut de Neurosciences de la Timone (ref 2016060618508941).

Alert monkey fMRI. The monkeys were first scanned for identifying Temporal Voice Areas (TVAs). All the details about the fMRI procedures are reported in ref. 3 (in which the two monkeys were called M2 and M3). Here, we give only a brief description of these details.

Functional scanning was done using an event-related paradigm with clustered-sparse acquisitions on a 3-Tesla MRI scanner (Prisma, Siemens Healthcare), equipped with an 8-channel surface coil (KU, Leuven). Ferrous oxide contrast agent (monocrystalline iron oxide nanoparticle, MION) was used for all the scanning sessions. Monkeys were trained to stay still in the scanner for a fixed period of 8 s to receive the reward. To avoid interference between sound stimulation and scanner noise, the scanner stopped acquisitions such that three repetitions of one of 96 stimuli (see below; interstimulus interval of 250 ms) were played on a silent background. Then, MION functional volumes were acquired using EPI sequences (multiband acceleration factor: 2, TR = 0.955 s). The analysis included 67 MION runs of M1 and 64 MION runs of M2. Voice-selective areas were identified as those regions responding significantly more to conspecific (macaque) vocalizations vs. nonvocal sounds (Fig. 1*A*).

fMRI-guided electrophysiology. Monkeys were chronically implanted with high-density microelectrode arrays (CerePort Utah Array, Blackrock Microsystems) to record extracellular activity in the fMRI-localized TVAs. Functional maps projected on the individual anatomical surfaces were used to calculate the exact position of the arrays. M1 was implanted with two 32-channel arrays in the anterior TVA (aTVA) of the right rostral superior temporal gyrus (rSTG; parcellation from the D99 macaque brain template; Ts2 from the AC map macaque brain template) and one 32-channel array in the right frontal cortex. M2 was implanted with three 32-channel arrays in the left rSTG (Ts2), of which one in the aTVA. In this study, we analyzed only data collected from the arrays implanted in the aTVA (i.e., two arrays for M1 and one array for M2; Fig. 1*A*).

Electrical signals were amplified and processed using a RZ2 BioAmp Processor (Tucker-Davis Technologies, Alachua, FL, USA) and sampled at 24,414 Hz. Raw data collected during recordings were high-pass filtered (300 to 5,000 Hz),

and spike sorting was performed offline using a fully automatic algorithm (MountainSort v4 0.2.3 and dependencies) (19). This algorithm detects, for each channel, individual clusters that are then filtered by applying specific thresholds to their quality parameters (firing rate > 0.5 Hz; noise overlap < 0.1; isolation > 0.95; signal-to-noise ratio > 2) to identify single units (for details about the parameters computed by the algorithm, see ref. 19). Finally, the mean waveforms of the remaining clusters were visually inspected to exclude neurons with irregular shapes.

In the main experiment, to limit the possible inclusion of the same neuron across sessions, we selected sessions separated in time by at least 3 d [M1: 7 ± 3 (mean ± SD) days' intervals between sessions on average; M2: 10 ± 13 d' intervals on average]. Our final dataset was composed of 353 single neurons, with 260 neurons from M1 across 9 recording sessions and 93 neurons from M2 across 5 recording sessions. Data for the control experiment come from 4 recording sessions of M1, collected 24 mo after those of the main experiment.

Auditory stimuli. The set of stimuli used for the main electrophysiological experiment was the same of that used in our previous fMRI study (3). Ninety-six acoustic stimuli from four main categories (i.e., human voices, macaque vocalizations, marmoset vocalizations, and nonvocal sounds) were played. Each category contained 24 stimuli divided into 4 subcategories of 6 stimuli (Fig. 1*C*). Specifically, human voices contained n = 6 female speech and n = 6 male speech [sentence segments from the set of stimuli used in the study of Moerel et al. (20)] and n = 6 female nonspeech and n = 6 male nonspeech [vocal-affect bursts selected from the Montreal Affective Voices dataset (21)]. Macaque vocalizations (kindly provided by Marc Hauser) included different call types: n = 6 coos, n = 6 grunts, n = 6 barks, and n = 6 screams. Marmoset vocalizations (kindly provided by Asif A. Ghazanfar) were also divided into different call types: n = 6 trills, n = 6 phees, n = 6 twitters, and n = 6 tsiks. Nonvocal sounds included both natural (n = 6 living and n = 6 nonliving) and artificial sounds, i.e., human actions (n = 6 artificial leaving) or not (n = 6 artificial nonleaving), from previous studies from our group (1, 22) or kindly provided by Petkov et al. (2) and Moerel et al. (20).

In the control experiment, we randomly interspersed the 96 original stimuli used in the main experiment with 96 spectrally matched (SM) and 96 temporally matched (TM) control stimuli, created from each original stimulus (Fig. 5*A*). The SM versions of the original stimuli consisted of a Gaussian white noise of the duration equals to the one of the original stimuli and filtered with the long-term spectrum of the original stimulus. TM stimuli consisted of a Gaussian white noise with the temporal envelope of the original stimuli. The temporal envelope was computed using the Hilbert transform modulus of the original stimuli. Indeed, some stimuli (e.g., macaque barks or grunts) are essentially defined by their temporal envelope.

In both experiments, stimuli were adjusted in duration so that all of them lasted 500 ms; they were resampled at 48,828 Hz and normalized by rms amplitude. Finally, a 10 ms cosine ramp was applied to the onset and offset of the stimuli.

Tonotopic organization of aTVA. Tonotopic organization of the recorded areas was tested by presenting band-passed noise stimuli (with central frequencies ranging from 125 Hz to 16,000 Hz in 1.75 octave steps; band weight: ⅓ octave) in alternative (1-d distant) sessions from those of the main experiment. For each recording site, we found the sound frequency eliciting the maximal response in the multiunit activity (MUA; band-pass filter: 300 to 5,000 Hz) computed over the first 200 ms after the sound onset. MUA was extracted from each recording site using the method previously described in refs. 23 and 24. We also tested frequency tuning at the single-unit level by presenting the band-pass noise stimuli in the control experiment.

Experimental setup and behavioral task. All recordings were performed in an acoustically insulated room. The monkeys sat in a primate chair with the head fixed by a noninvasive modular restriction mask (MRM) developed in our laboratory. Auditory stimuli were presented through a RZ6 Multi-I/O Processor (Tucker-Davis Technologies) and transduced by two 8,020 Genelec speakers, which were positioned at ear level 72 cm from the head and 60 degrees to the left and right. Stimuli were delivered at a sound pressure level of approximately 92 dB. Hand detection was achieved using two optical sensors.

Monkeys were trained to perform a pure tone detection task (Fig. 1*B*). They were required to hold a bar with both hands for 1,500 to 2,000 ms to trigger the presentation of sounds. In each trial, from three to seven stimuli (interstimulus interval: 285 to 542 ms) were played after which a 500 ms 1,000-Hz pure tone was presented. The pure tone instructed the monkeys to release the bar to receive the juice reward (correct trials). If the monkeys released the bar before the pure tone presentation (false alarm trials) or did not release the bar (miss trials; upper reaction time: 250 ms), no reward was given. The stimuli were randomly presented, but two sounds of the same category were never played one after the other. We presented all the stimuli before their repetition, allowing a similar number of repetitions played.

Electrophysiological Data Analysis. The database for this study comes from recording sessions in which both monkeys successfully performed the task with an average performance of 84 ± 7% (mean ± SD) of correct trials for M1 and 74 ± 7% of correct trials for M2. In the main experiment, each of the 96 stimuli was repeated on average 18 ± 2 times in M1 sessions and 16 ± 3 times in M2 sessions (resulting in 1,725 ± 146 sounds played on average for M1 and 1,497 ± 304 for M2).

The analyses were performed using MATLAB software (The MathWorks, Inc.) and open source statistical software R. The spike times were binned at a resolution of 1 ms. Peristimulus time histograms (PSTHs) were smoothed with a Gaussian kernel of 10 ms. We included all the trials in the neuronal analyses given the low percentage of wrong (false alarm and miss) trials. Moreover, no significant difference (Wilcoxon rank-sum test; all *P*'s < 0.05/96 after Bonferroni correction) was observed in the neuronal activity (averaged across the stimulus presentation time) of correct vs. all trials in response to each sound in each neuron of M1 and M2.

The baseline activity was defined as the average firing rate (FR) during 100 ms preceding the stimulus onset. The analyzed period of the task corresponded to the stimulus presentation time, i.e., the 500 ms window following the stimulus onset. To compute z-scores, the average response to each stimulus was normalized to SD units with respect to the baseline.

The maximum response time of each neuron was estimated as a 150 ms window centered on its peak of activity computed across all sounds. To depict neurons' responsivity to each of the 96 stimuli, we computed for each neuron the z-scored FR during the maximum response time (Fig. 1*D*).

Auditory and category responsiveness. Auditory responsiveness of neurons was evaluated by computing z-scored responses to each stimulus every 10 ms bins during the stimulus presentation time. Neurons were considered auditory responsive if the z-score in response to one or more stimuli was above 2.5 SDs for at least three consecutive bins, starting at a time point less than 300 ms post-stimulus onset (approach similar to ref. 12). Only auditory responsive neurons were further analyzed.

In the control sessions, auditory responsiveness was computed as in the main experiment but using the z-scored responses to each stimulus subcategory (because of the too few repetitions of each stimulus) and taking a threshold of 2 SDs.

The preference of neurons (i.e., higher FRs response) for one sound category (human, macaque, marmoset, and nonvocal) was evaluated by comparing FRs (less the baseline) of the four categories during the maximum response time (one-way ANOVA, factor: category). For neurons showing a main effect (*P*-value threshold of 0.05), we defined the preferred category as the one eliciting the maximal response.

Selectivity of neurons. Voice-selective neurons were classified using the Voice Selectivity Index (VSI) criterion as previously done in auditory studies (12) and according to the Face Selectivity Index used in visual studies (13). We computed two VSIs contrasting human voices vs. nonvocal sounds (Hum VSI) and macaque vocalizations vs. nonvocal sounds (Mac VSI). The VSI was defined as follows:

$$\mathrm{VSI} = \frac{\mathrm{mean}_{\mathrm{voice}} - \mathrm{mean}_{\mathrm{non-vocal}}}{\mathrm{mean}_{\mathrm{voice}} + \mathrm{mean}_{\mathrm{non-vocal}}},$$

where $\mathrm{mean}_{\mathrm{voice}}$ is the average FR for human or macaque category during the maximum response time, and $\mathrm{mean}_{\mathrm{non-vocal}}$ is the average FR for nonvocal category during the maximum response time.

A VSI of 0 indicates equal responses to voices and nonvocal sounds. A VSI of 0.33 indicates twice as strong a response to voices as to nonvocal sounds. Conversely, a VSI of −0.33 indicates twice as strong a response to nonvocal sounds as to voices. We used these thresholds to establish the voice/nonvocal selectivity of neurons as previously done (12, 13). For cases where $\mathrm{mean}_{\mathrm{voice}} > 0$ and $\mathrm{mean}_{\mathrm{non-vocal}} < 0$, VSI was set to 1; for cases where $\mathrm{mean}_{\mathrm{voice}} < 0$ and

$\text{mean}_{\text{non-vocal}} > 0$, VSI was set to −1 (25). We also computed a SI contrasting human voices and macaque vocalizations (Hum-Mac SI) with the same method explained above.

Decoding analysis. We used a maximum correlation coefficient classifier [MCC; as implemented in the MATLAB neural decoding toolbox (26)] to analyze the aTVA neuronal population or the different subpopulations of neurons. The classifier was trained to discriminate either the four sound categories, either human or macaque vocalizations from the other categories, or, within each of these two categories, between subcategories of sound.

To train the classifier, trials (i.e., stimuli) were labeled based on their category/subcategory, and firing rates from trials and neurons were binned in 100 ms sliding windows, every 10 ms bins. We tested 56 consecutive bins, from −100 ms (i.e., a time window from −100 ms to 0 ms) to +450 ms from the stimulus onset. Note that these 100 ms bins are plotted such that the decoding accuracy is aligned to the center of each bin. For each bin, a different classifier was trained/tested.

Z-score normalization was applied to each neuron to give equal weight to all the units regardless of firing rate. Then, the classifier was trained using a k cross-validation splits procedure: The classifier was trained using k−1 splits and then tested on the remaining split. Particularly, for classification between categories, we used 100 cross-validation splits, and for classification between subcategories, we used 25 cross-validation splits. All possible train/test splits were tested, and this process was repeated 50 times (i.e., 50 runs) with different subsets of trials. The classification accuracy from these runs was then averaged.

To assess whether the obtained decoding accuracies were above chance, we ran a permutation test that consisted of repeating the full decoding procedure 100 times with the labels of categories/subcategories randomly shuffled. We obtained a null distribution of shuffled data, and the decoding results were considered significantly above chance if they were greater than all the shuffled data in the null distribution [*P*-value threshold of $P = 1/(100 * 56) = 0.0002$]. The latency of when the *P*-values are first above chance corresponds to the first time bin of three consecutive bins with *P*-values below the *P*-value threshold.

To determine the similarity between two classification accuracy distributions, we computed for each time bin a distribution-free overlapping index (η) using the overlapping package for R (27). The overlapping index η represents the proportion of the overlapping area between the probability density functions of two distributions. In this sense, an overlapping index of $\eta(A,B) = 0$ indicates that f _ A (X) and f _ B (X) are distinct. Two distributions were considered as significantly different for $\eta < 0.05$ for at least three consecutive time bins.

Representational similarity analysis. We conducted a representational similarity analysis (RSA) as it is a powerful tool for comparing representational geometries for a given stimulus set across species and measurement techniques (14, 15). We generated Representational Dissimilarity Matrices (RDMs) time series from human- and macaque-selective subpopulations. For each of the 16 subcategories, we pulled out a vector of z-scored average firing rate from the neurons, every 10 ms bins (±50 ms sliding windows). 16 × 16 RDMs were generated for each bin by computing for each pair of subcategories the Euclidean distance between their population vectors. In total, we obtained 51 neuronal RDMs for both the human-selective and the macaque-selective subpopulations.

For each subpopulation of neurons, we compared the 51 neuronal RDMs with four categorical model RDMs (human, macaque, marmoset, and nonvocal) as previously done in ref. 3. Model RDMs were four theoretical patterns of pairwise dissimilarities in our stimulus set representing the ideal distinction between one specific category and the others (Fig. 3 *B*, *Bottom*). Planned comparisons between each of the 51 neuronal RDMs and model RDMs were performed by comparing the between vs. the within portions of the neuronal RDM predicted by each model (*SI Appendix*, Fig. S6) using bootstrapped two-sample *t* tests (10,000 iterations, one-tailed), with Bonferroni correction for multiple comparisons, resulting in a corrected *P*-value threshold of $P = 0.05/(51*4) = 2.451\text{e-}04$.

We also compared human- and macaque-selective neuronal RDMs within the human and within the macaque categories using a Mann–Whitney–Wilcoxon test (*P*-value threshold of 0.05).

For all comparisons, the latency of when the *P*-values are first significant corresponds to the first time bin of three consecutive bins with *P*-values below the corrected *P*-value threshold.

The macaque- and human-selective neuronal RDMs, computed at the time of peak association with the respective model RDM, were also compared to the categorical model RDMs, to RDMs computed from fMRI data, and to acoustical measures of the stimuli (previously shown in ref. 3).

In this analysis, categorical model RDMs represented ideal categorical distinction between one specific category vs the others using binary values of 1 as maximal pairwise distance and 0 as minimal pairwise distance.

fMRI RDMs represented the Euclidean distance between the 16 subcategories in multivoxel activity space of the aTVA of humans (averaged across the left and the right hemispheres of six subjects) and macaques (averaged across the right hemisphere of M1 and the left hemisphere of M2, i.e., the brain areas where the arrays were implanted).

Acoustical RDMs were generated for each of three measures–loudness, spectral center of gravity (SCG), and pitch–by computing for each pair of stimulus subcategories the difference of the measure averaged across stimuli of each subcategory. Loudness and SCG (an acoustical correlate of timbre brightness) were estimated by modeling each sound using the time-varying loudness model by Glasberg and Moore (28). Pitch was estimated by modeling each sound using the Y IN pitch extraction model by de Cheveigne and Kawahara (29).

The comparison between the human- and the macaque-selective neuronal RDMs, and between each of these with the fMRI, the acoustical and the categorical model RDMs was run through a bootstrapped Spearman's rho correlation (10,000 iterations), with Bonferroni correction for multiple comparisons resulting in a corrected *P*-value threshold of $P = 0.05/17 = 0.0029$.

Visual representation of the pattern of correlation between RDMs in Fig. 4*C* was obtained via a multidimensional scaling (MDS) arrangement reflecting the dissimilarity structure of RDMs using the RSA toolbox (30).

Data, Materials, and Software Availability. The preprocessed data for the main experiment in this paper are available in the Zenodo repository: https://doi.org/10.5281/zenodo.11284105 (31).

ACKNOWLEDGMENTS. This work was funded by Fondation pour la Recherche Medicale AJE201214 (P.B.); Agence Nationale de la Recherche grants ANR-16-CE37-0011-01 (PRIMAVOICE) (P.B.) ANR-16-CONV-0002 (Institute for Language, Communication and the Brain) (P.B.) and ANR-11-LABX-0036 (Brain and Language Research Institute) (P.B.); Excellence Initiative of Aix-Marseille University (A*MIDEX) (P.B.); and European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 788240) (P.B.).

1. P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, B. Pike, Voice-selective areas in human auditory cortex. *Nature* **403**, 309–312 (2000).
2. C. I. Petkov *et al.*, A voice region in the monkey brain. *Nat. Neurosci.* **11**, 367–374 (2008).
3. C. Bodin *et al.*, Functionally homologous representation of vocalizations in the auditory cortex of humans and macaques. *Curr. Biol.* **31**, 4839–4844.e4 (2021).
4. S. Sadagopan, N. Z. Temiz-Karayol, H. U. Voss, High-field functional magnetic resonance imaging of vocalization processing in marmosets. *Sci. Rep.* **5**, 10950 (2015).
5. A. Jafari *et al.*, A vocalization-processing network in marmosets. *Cell Rep.* **42**, 112526 (2023).
6. A. Andics, M. Gácsi, T. Faragó, A. Kis, Á. Miklósi, Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. *Curr. Biol.* **24**, 574–578 (2014).
7. A. J. Hall, B. E. Butler, S. G. Lomber, The cat's meow: A high-field fMRI assessment of cortical activity in response to vocalizations and complex auditory stimuli. *NeuroImage.* **127**, 44–57 (2016).
8. C. Bodin, P. Belin, Exploring the cerebral substrate of voice perception in primate brains. *Philos. Trans. R. Soc. B Biol. Sci.* **375**, 20180386 (2020).
9. J. K. Hesse, D. Y. Tsao, The macaque face patch system: A turtle's underbelly for the brain. *Nat. Rev. Neurosci.* **21**, 695–716 (2020).
10. A. Andics, T. Faragó, "Voice perception across species" in *The Oxford Handbook of Voice Perception*, S. Frühholz, P. Belin, Eds. (Oxford University Press, 2018).
11. M. J. Arcaro, P. F. Schade, J. L. Vincent, C. R. Ponce, M. S. Livingstone, Seeing faces is necessary for face-domain formation. *Nat. Neurosci.* **20**, 1404–1412 (2017).
12. C. Perrodin, C. Kayser, N. K. Logothetis, C. I. Petkov, Voice cells in the primate temporal lobe. *Curr. Biol. CB.* **21**, 1408–1415 (2011).
13. D. Y. Tsao, W. A. Freiwald, R. B. H. Tootell, M. S. Livingstone, A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–674 (2006).
14. N. Kriegeskorte, Representational similarity analysis–connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008), 10.3389/neuro.06.004.2008.
15. N. Kriegeskorte *et al.*, Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron.* **60**, 1126–1141 (2008).
16. C. Perrodin, C. Kayser, N. K. Logothetis, C. I. Petkov, Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices. *J. Neurosci.* **34**, 2524–2537 (2014).
17. S. Dehaene, L. Cohen, Cultural recycling of cortical maps. *Neuron* **56**, 384–398 (2007).
18. I. Gauthier, M. J. Tarr, Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior. *J. Exp. Psychol. Hum. Percept. Perform.* **28**, 431–446 (2002).

19. J. E. Chung *et al.*, A fully automated approach to spike sorting. *Neuron.* **95**, 1381–1394.e6 (2017).
20. M. Moerel, F. De Martino, E. Formisano, Processing of natural sounds in human auditory cortex: Tonotopy, spectral tuning, and relation to voice sensitivity. *J. Neurosci.* **32**, 14205–14216 (2012).
21. P. Belin *et al.*, Human cerebral response to animal affective vocalizations. *Proc. R. Soc. B Biol. Sci.* **275**, 473–481 (2008).
22. A. Capilla, P. Belin, J. Gross, The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. *Cereb. Cortex.* **23**, 1388–1395 (2013).
23. M. Mattia *et al.*, Heterogeneous attractor cell assemblies for motor planning in premotor cortex. *J. Neurosci.* **33**, 11155–11168 (2013).
24. M. Giamundo *et al.*, Neuronal activity in the premotor cortex of monkeys reflects both cue salience and motivation for action generation and inhibition. *J. Neurosci.* **41**, 7591–7606 (2021), 10.1523/JNEUROSCI.0641-20.2021.
25. W. A. Freiwald, D. Y. Tsao, Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* **330**, 845–51 (2010), 10.1126/science.1194908.
26. E. M. Meyers, The neural decoding toolbox. *Front. Neuroinform.* **7**, 8 (2013), 10.3389/fninf.2013.00008.
27. M. Pastore, A. Calcagnì, Measuring distribution similarities between samples: A distribution-free overlapping index. *Front. Psychol.* **10**, 1089 (2019).
28. B. R. Glasberg, B. C. J. Moore, A model of loudness applicable to time-varying sounds. *J. Audio Eng. Soc.* **50**, 331–342 (2002).
29. A. de Cheveigne, H. Kawahara, YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* **111**, 1917–1930 (2002).
30. H. Nili *et al.*, A toolbox for representational similarity analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).
31. M. Giamundo *et al.*, Data from "A population of neurons selective for human voice in the monkey brain." Zenodo. https://doi.org/10.5281/zenodo.11284105. Deposited 24 May 2024.